СЁРЧ

Автоматическая система рекомендаций и подбора контента

Феофилактов Владислав, 11А класс, МБОУ СОШ №54
Научный руководитель: Савельева Елена Николаевна

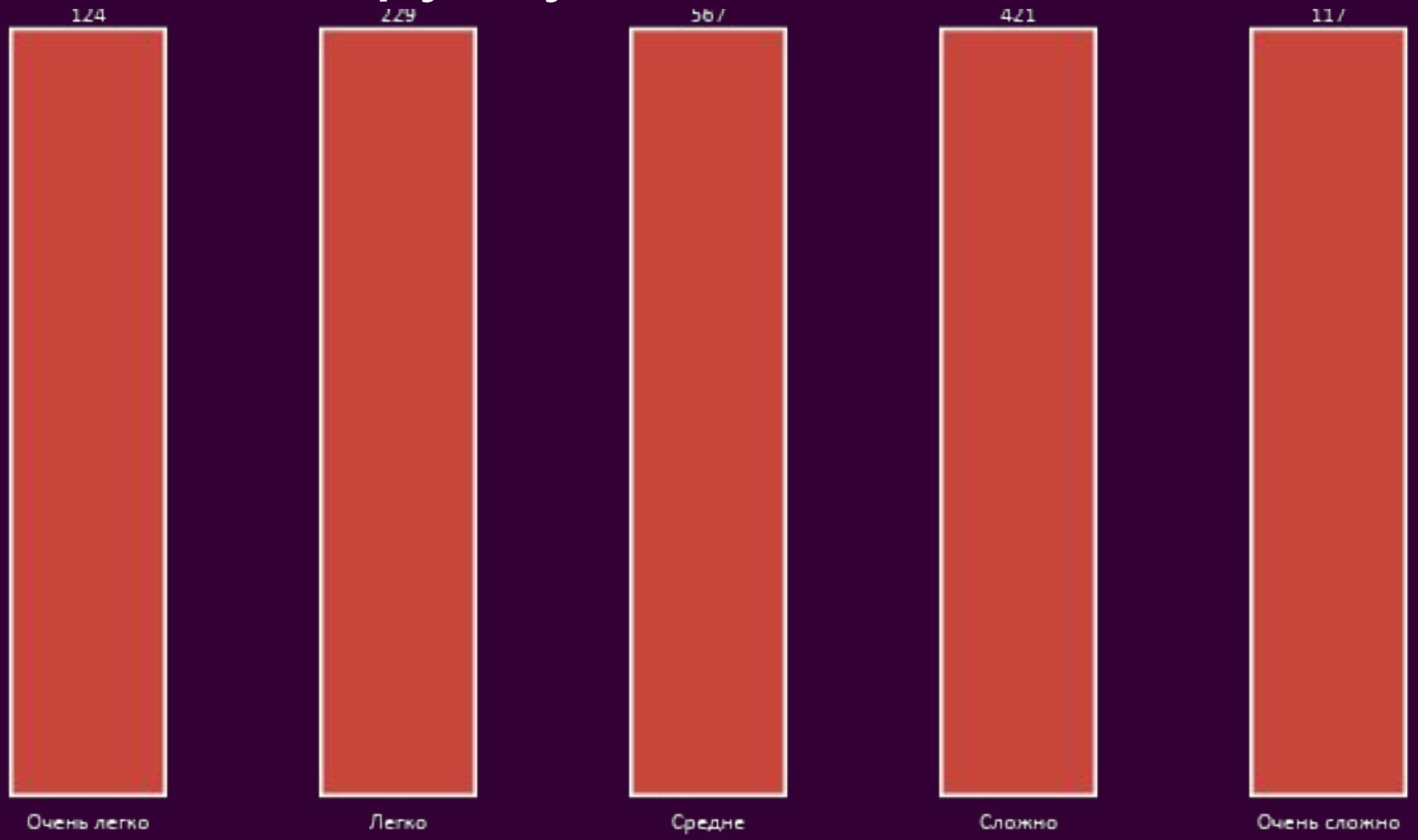
Постановка проблемы

В современном мире социальные сети играют важную роль.
Они позволяют пользователям получать любую информацию

Но ориентироваться в этом потоке почти

невозможно

Насколько Вам сложно найти интересную группу ВКонтакте?



Цели и задачи

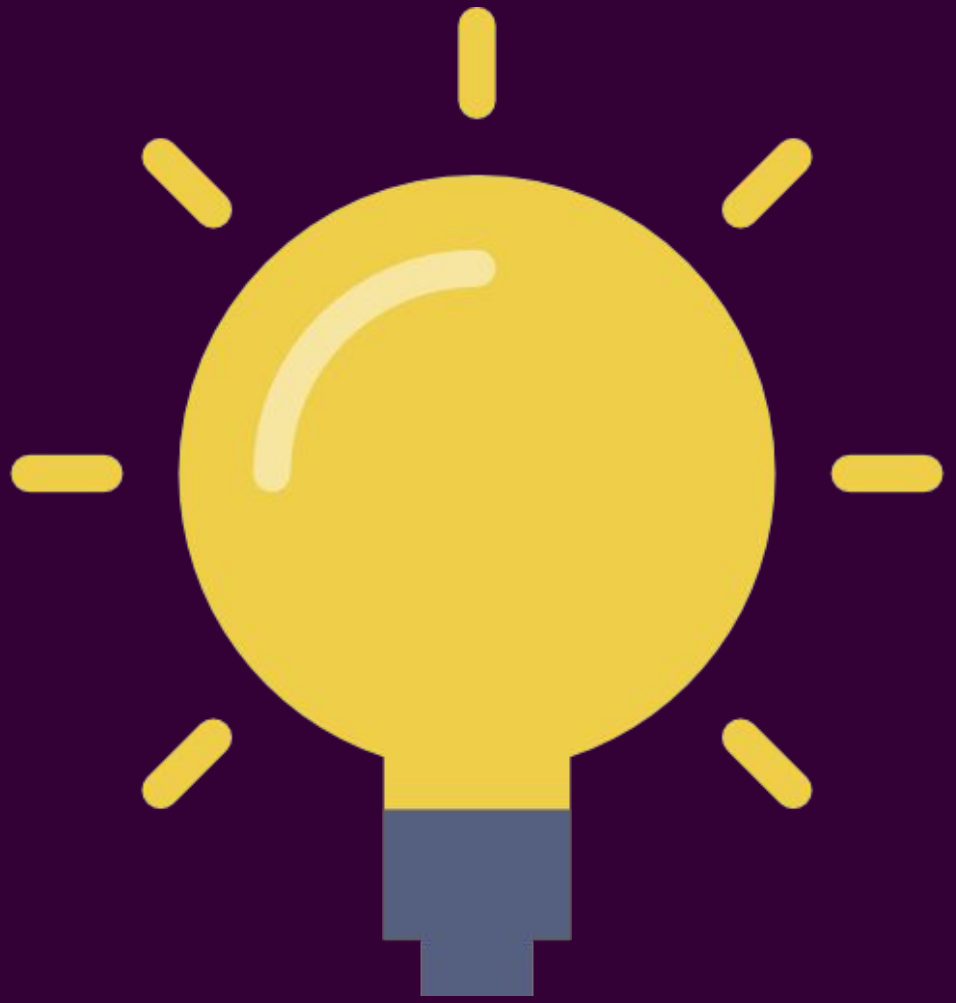
1. Создание сервиса, который поможет пользователям в поиске интересных сообществ ВКонтакте

2. Разработка набора алгоритмов машинного обучения и обработки больших данных, который решал бы задачу рекомендации контента пользователю

3. Разработка программной архитектуры сервиса и ее реализация

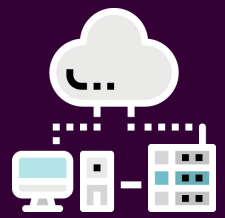
Гипотеза

Если получится создать сервис рекомендаций контента социальных сетей, то пользователям будет удобнее его получать



Краткое описание

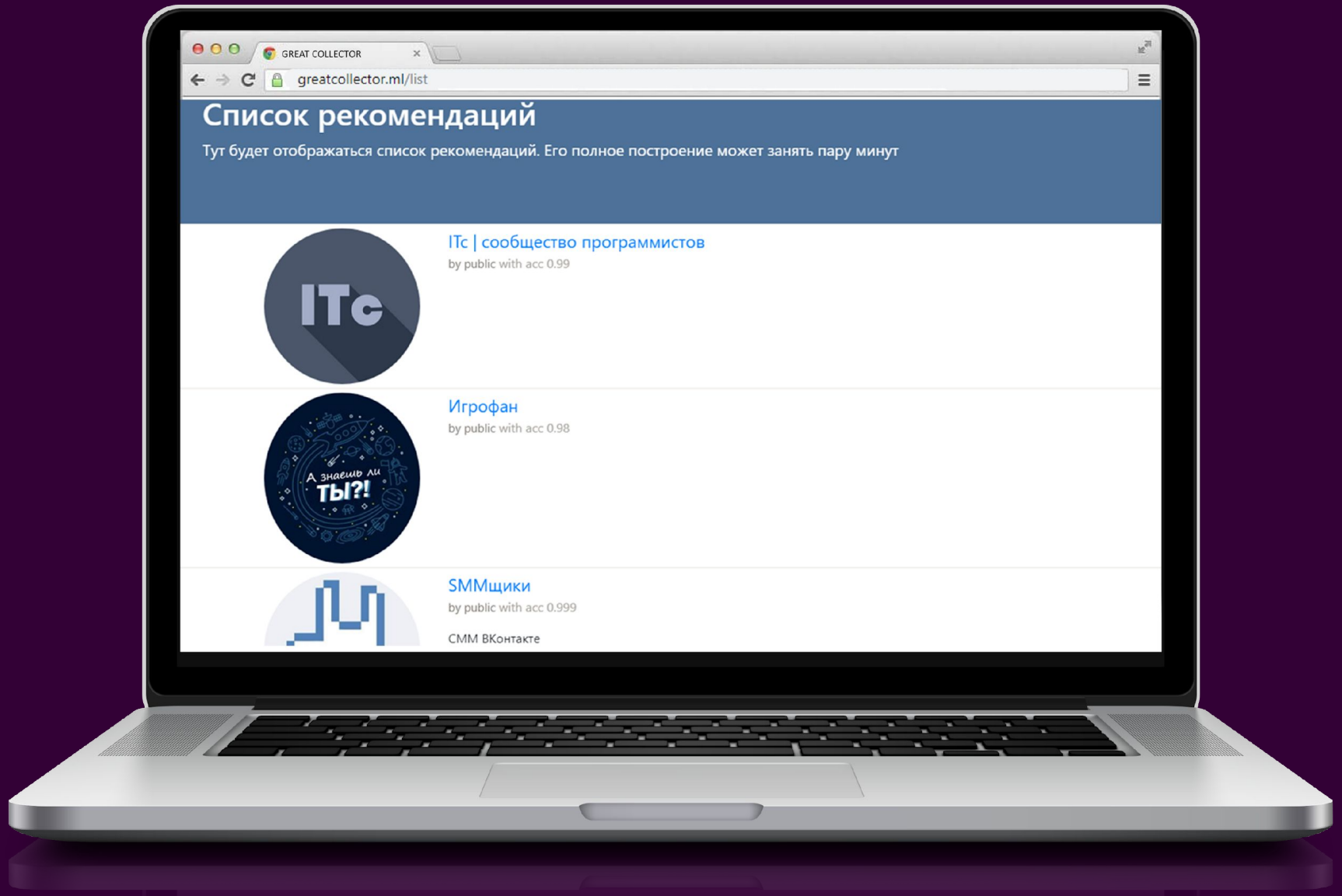
Сёрч – система подбора и рекомендаций сообществ ВКонтакте



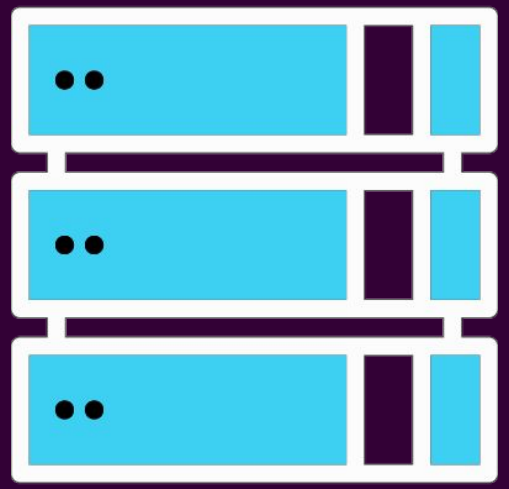
Работает на технологиях машинного обучения и обработки больших данных

Используются передовые технологии

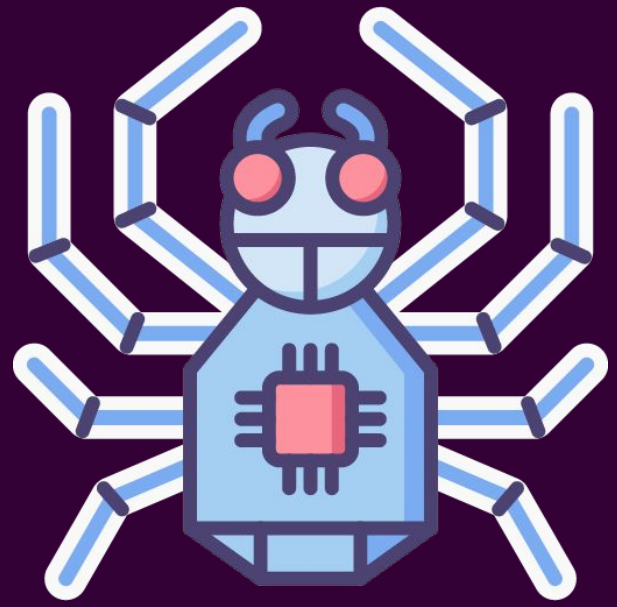
Пример выдачи результатов



Получение данных



Необходимо для пополнения базы данных новыми сообществами



Происходит при помощи поискового бота

Принципы работы

Нейрон – наименьший элемент нейронной сети

Функция нейрона

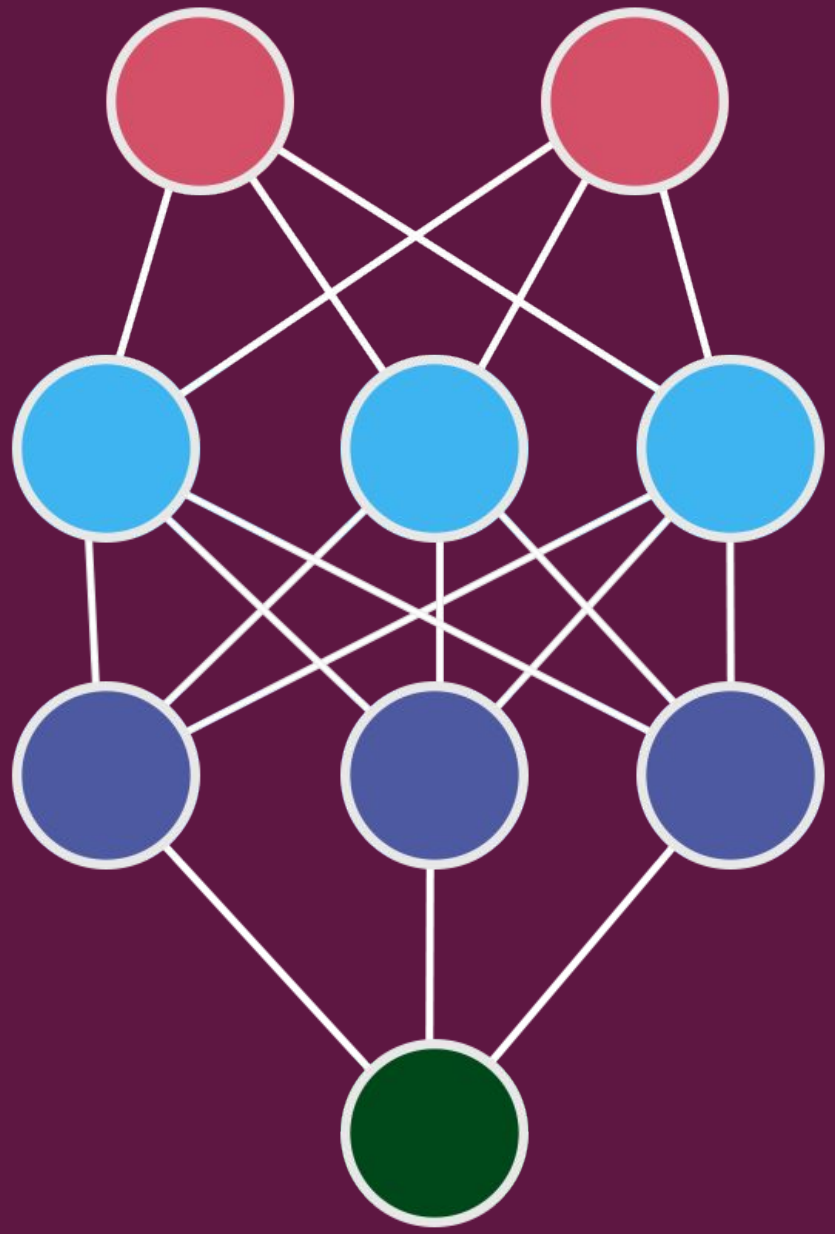
Суммирование сигналов

Нейронная сеть – множество нейронов, связанных в одно целое. Сейчас они используются в самых разных задачах (определение объектов по фото, распознавание речи, её синтез и прочее)

Функции нейронных сетей

Обучение

Обработка данных



Математический вектор

- Вектор – упорядоченное множество элементов
- Размерность вектора – количество элементов в нём

$$\vec{a}\{42; 0; 1\}$$

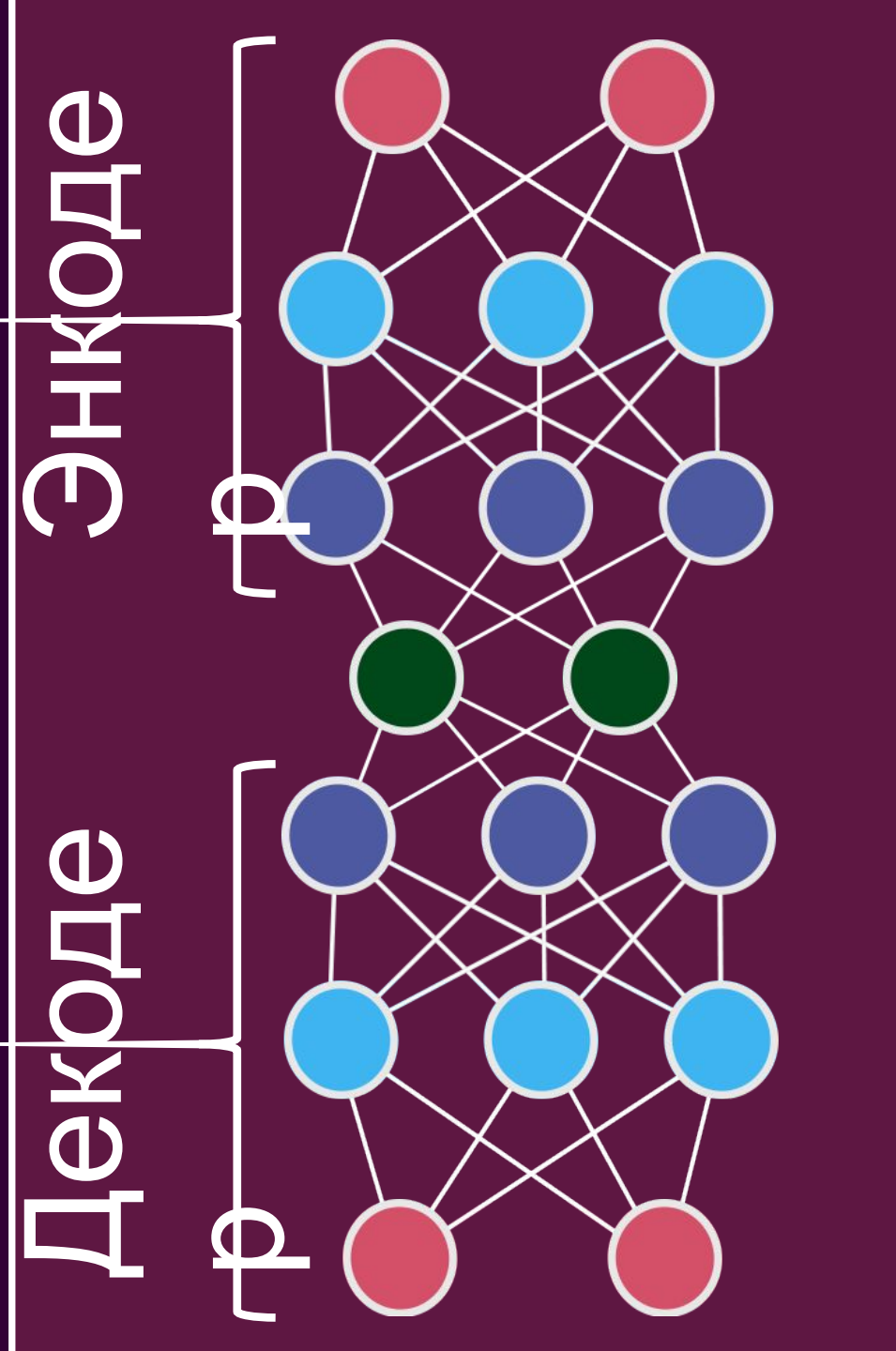
У вектора $\vec{a}$ размерность 3

Автоэнкодер

Энкодер получает на вход вектор и уменьшает его размерность. На выходе получается внутреннее представление

Декодер на вход получает от энкодера внутреннее представление входных данных и пытается по ним восстановить сами входные данные

Декодер используется только при обучении нейронной сети



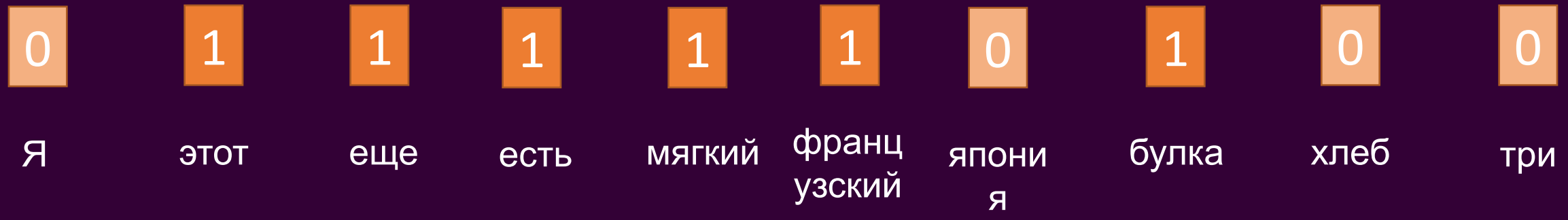
Предобработка текста

1. Удаляются небуквенные символы
2. В словах исправляются ошибки
3. Они принимают начальную форму, маленькие слова отбрасываются

0. Съешь же ещё этих мяхких французких булок да😊выпей чаю!
1. Съешь же ещё этих мяхких французких булок да выпей чаю
2. Съешь же еще этих мягких французских булок да выпей чаю
3. Есть ещё этот мягкий французский булка пить чай

Словарь

Есть ещё этот мягкий французский булка



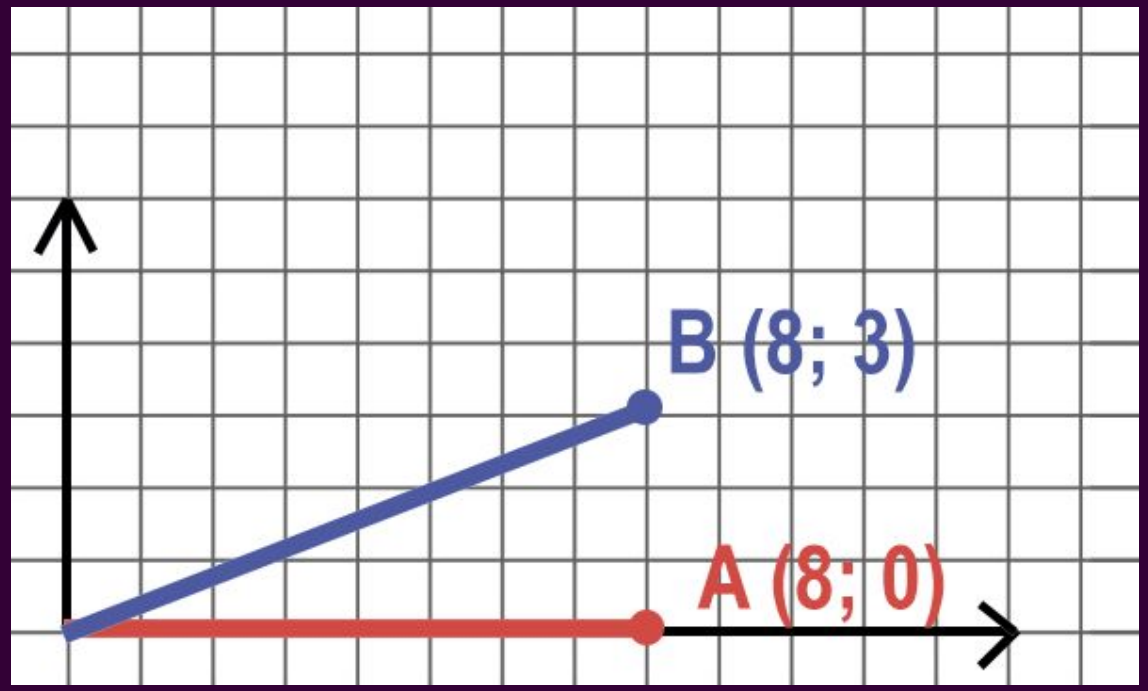
Что дальше?

1. Получили много текстовых данных (поисковый бот)
2. Получили их векторное представление (энкодер)
3. ???

Сходство векторов

- Сходством векторов является косинус угла между векторами.

Пусть есть два вектора - $\vec{a}\{8; 0\}$ и $\vec{b}\{8; 3\}$. Обозначим их на плоскости:



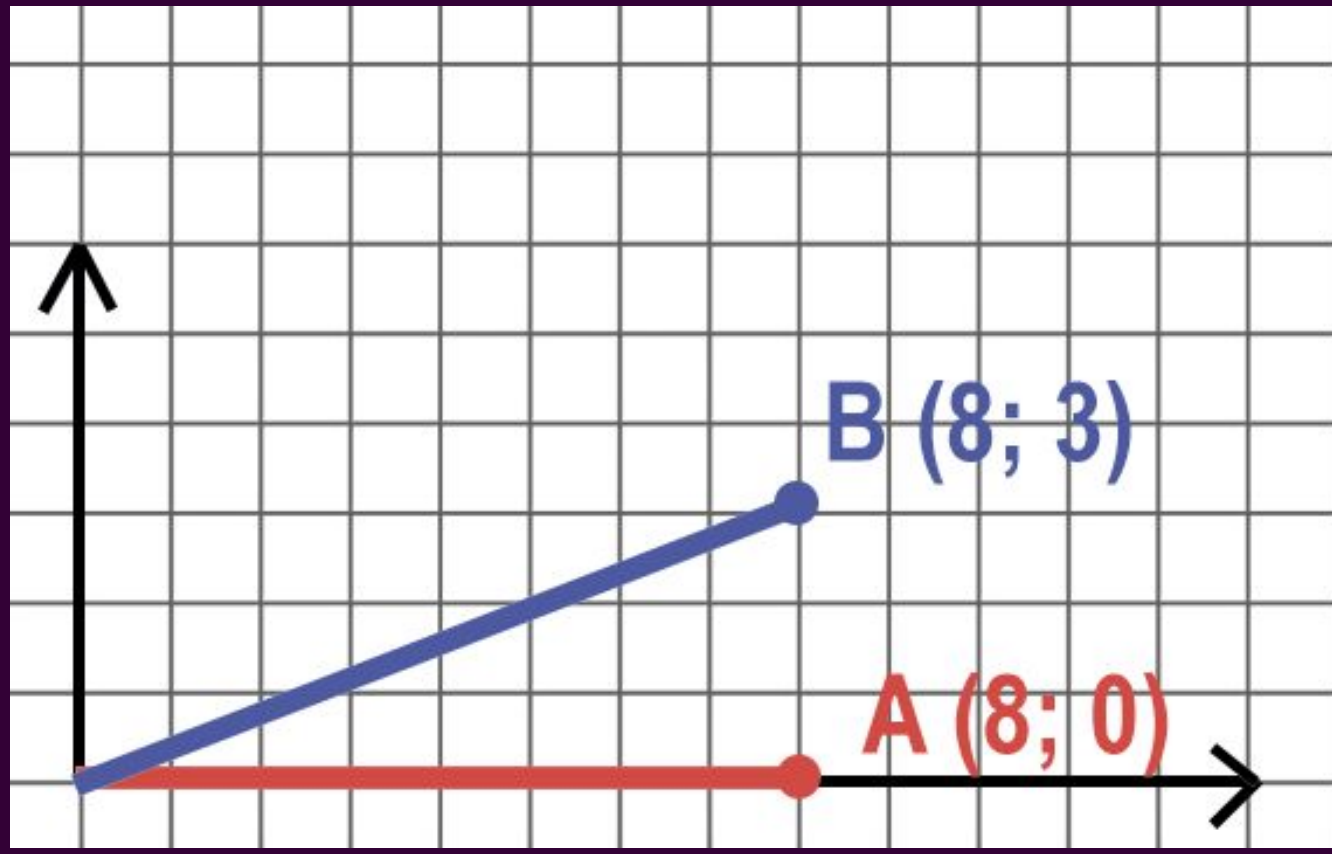
Тогда угол между этими векторами находится по формуле:

$$\cos \widehat{\vec{a}\,\vec{b}} = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

У нас эта формула принимает вид:

$$\cos \widehat{\vec{a}\,\vec{b}} = \frac{8 * 8 + 0 * 3}{\sqrt{8^2 + 3^2} * \sqrt{8^2 + 0^2}} =$$

$$= \frac{64}{\sqrt{73 * 64}} \approx 0.94$$



Сходство наших векторов примерно равно 0.94. Чем больше значение – тем выше сходство

Получение общего вектора сообщества

- Общий вектор сообщества – такой вектор, который характеризует все посты в рамках одного сообщества

- Получается методом сложения всех векторов постов сообщества и делением полученного вектора на их количество

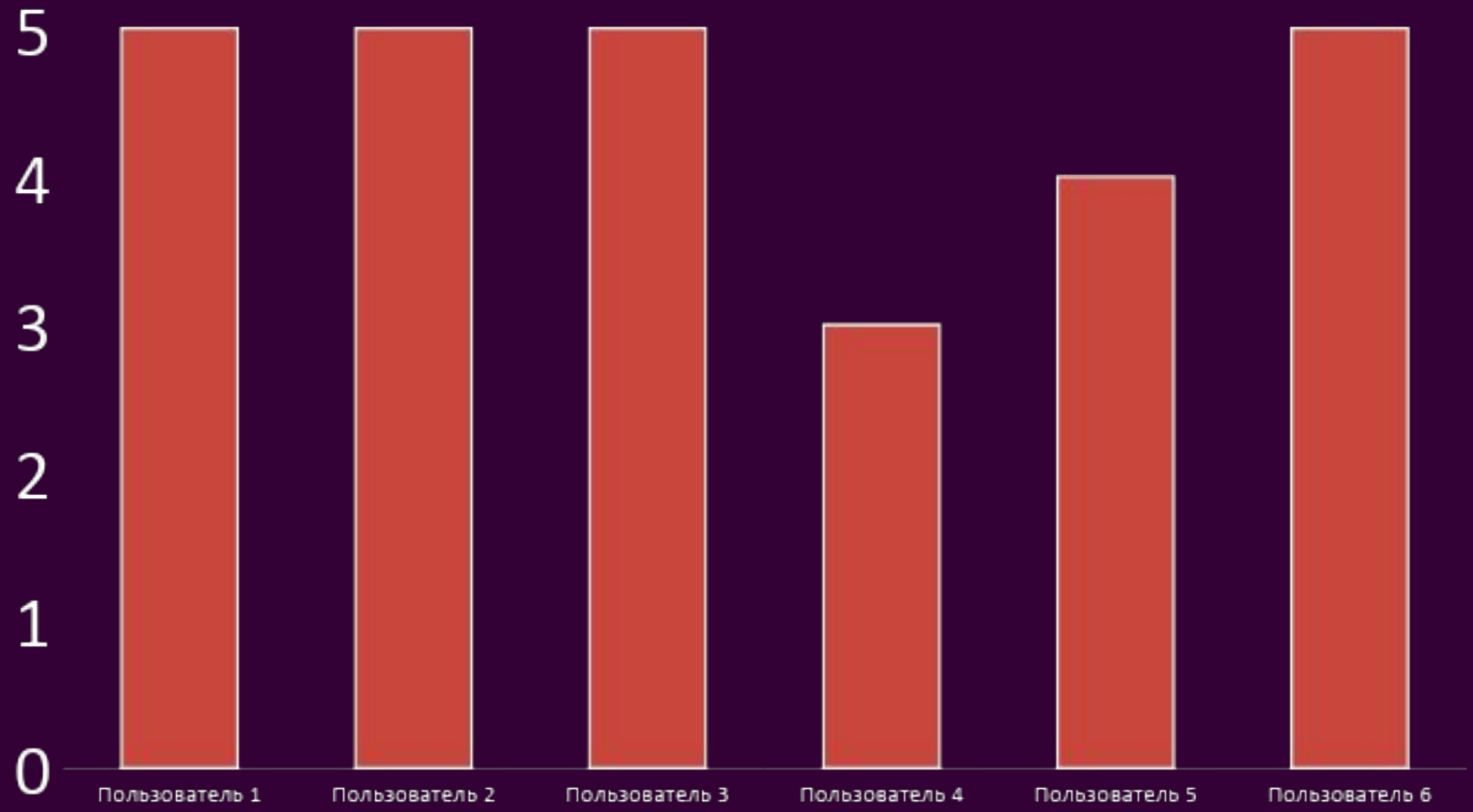
- $\vec{a}\{42; 1; 7\}$, $\vec{b}\{42; 2; 2\}$, $\vec{c}\{39; 3; 6\}$.

Общий вектор $\vec{n}\{42 + 42 + 39;\ 1 + 2 + 3;\ 7 + 2 + 6\} * \frac{1}{3}$,

$\vec{n}\{41;\ 2;\ 5\}$

Общая архитектура реализации

Оценки пользователей-тестеров



Итоги

- Проект, который поможет пользователям в поисках информации, который активно разрабатывается и будет готов *летом*
- Произведены различные исследования на темы использования различных алгоритмов, способов обработки, хранения информации и прочие
- Изучены и разработаны новые технологии и способы применения уже имеющихся
- Разработана архитектура проекта

Выводы

- Нейронные сети стали неотъемлемой частью современной IT индустрии
- Они позволяют делать многие интересные вещи, в том числе и обрабатывать текстовую информацию
- Интернет-сообществу можно помочь при помощи создания сервисов подбора и рекомендации контента, а значит, можно помочь и пользователям в самообразовании

Цели на будущее

- Найти средства на несколько мощных серверов
- Реализовать сравнение сообществ не только по тексту постов, но и по другим параметрам (фото и видео, другие медиа)
- Разработать алгоритмы создания еженедельных подборок сообществ по тематикам на главной странице
- Оптимизировать все процессы получения и обработки информации
- Создать красивый и удобный дизайн сайта
- Привлечь пользователей

Спасибо за внимание

Группа проекта
ВКонтакте

Литература

1. https://habr.com/ru/post/436636/ - Моя статья на habr.com
2. https://habr.com/ru/post/331382/ - Информация про автоэнкодеры
3. https://habr.com/ru/post/123671/ - Цикл статей про интернет-поисковик (идея хранения большого объёма информации)
4. https://habr.com/ru/post/312450/ - Вводный курс статей про нейронные сети
5. https://vk.cc/99TFSR [англ.] – простое описание алгоритма сравнения векторов